Rise of Clean Energy: Comprehensive Analysis of Renewable Energy Production across Europe

Introduction

The global energy landscape has drastically changed since Russia's invasion of Ukraine and the rise of environmental sustainability goals. Specifically for countries within the EU, who primarily depended on Russia's fossil fuel imports, clean energy transitions to renewable sources have accelerated significantly. Hence, our project focuses on analyzing a comprehensive dataset on renewable energy plants across Europe to derive insights about the continent's transition to cleaner energy. By exploring geographical distributions, technological trends, and other features within the dataset we provide institutions at stake the appropriate insights for strategic decision-making.

Prior to the exploration of the dataset, we hypothesize that there are regions within the EU where renewable energy development is attractive and currently underserved. Therefore, the project aims to provide insights for strategic site selection, forecasting future energy demand and trends, and market entry strategy for renewable energy companies. The insights from the project enhance the likelihood of expansion for renewable energy companies by highlighting markets with high growth potential, a supportive regulatory environment, and understanding local demand. With the application of our insights, the overhead costs of renewable energy investments are minimized, as our solutions abide by markets with high yield potential and favorable conditions.

The primary questions that guided our analysis throughout the project include:

- What is the distribution of renewable energy power plants across different countries in Europe? How about concentration of plants by location?
- How has the installed electrical capacity evolved over the years?
- What are the adoption trends of different renewable energy technologies in different countries?

Data Description

There are 11 files with a total of 4,125,218 rows ranging between 10 and 2,115,921 rows with qualitative and quantitative columns like location, energy source, commissioning date, electrical capacity, etc. The files are large and given a team of analysts, it is not ideal for each analyst to load the files into the computer memory. The data lacks more insightful features like number of energy source users, details about the regulatory policies and pricing and costs of setting up plants for the renewable energy plant, these would have been very beneficial for the analysis.

Data Interaction

In the context of our project, Presto (ANSI SQL dialect) and Hive (HQL dialect) work in parallel to provide analysts to interact with the data. To illustrate, Presto is favored for interactive and ad-hoc querying, while Hive is optimized for batch processing and ETL workflows. Presto can connect to multiple meta stores/catalogs, easier to set up a Python connection, and Presto does not need to first initialize a session to run unlike Hive. Furthermore, Presto can not drop none empty Hive tables or schemas so we prevent accidental table drops. More about this is summarized by Natarajan [1] and Ecuba [2]. For our analysis later in the project, we employ Presto to interact with the data.

Data Wrangling

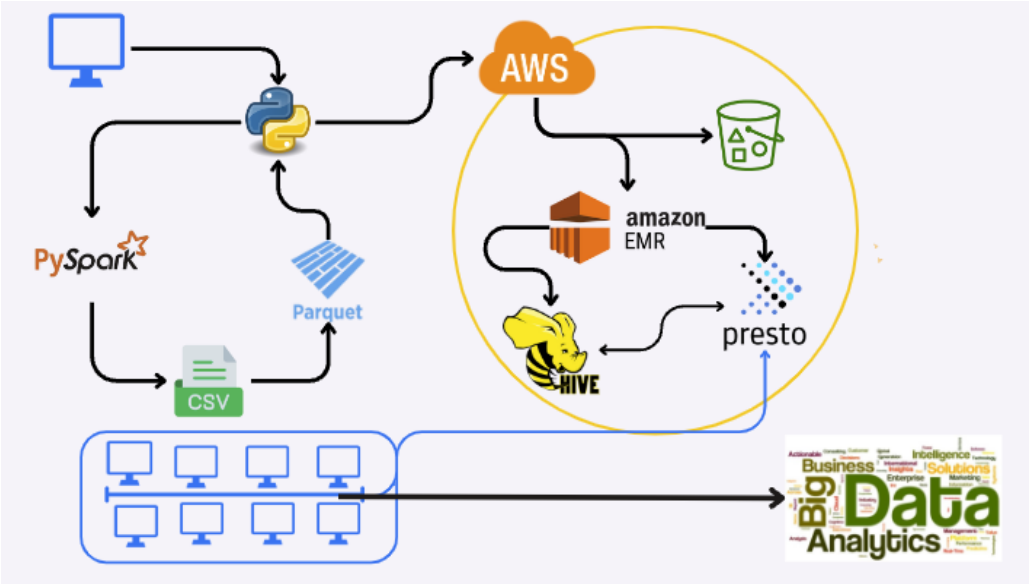


Figure 1: Project Data Model

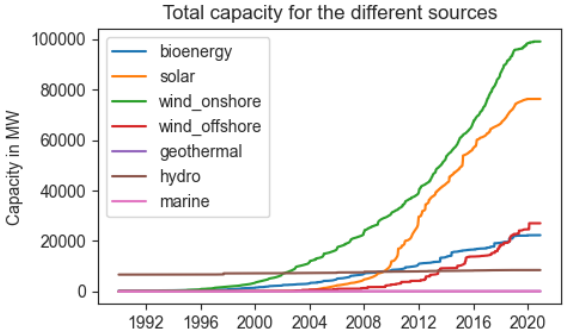
Technical Use Case

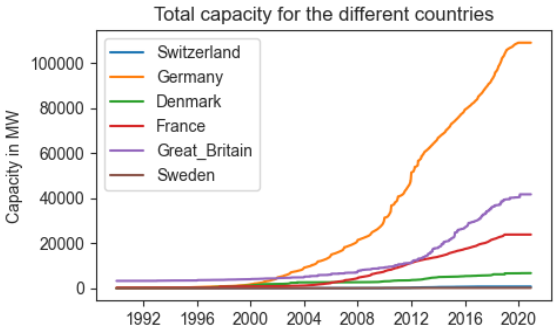
In our project, we have designed a data pipeline that centralizes the data ingestion process across AWS into Python. The data generated by renewable energy sources across different European countries is massive, hence we have optimized the process of making the data readily available for a team of analysts. Our solution primarily utilizes Python's Boto3 library to interact with the AWS Console to perform the following tasks:

1. Create an S3 bucket and create an EMR cluster with Hadoop, Hive, Hue and Presto and other configurations specified within a JSON file.
2. Read the files using PySpark, keeping a copy of each file's schema, then convert them to Parquets for storage on a folder in S3.
3. Once the EMR cluster is created and ready for use, copy the Parquets from S3 bucket to HDFS folder. Additionally, the S3 bucket is deleted upon completion of the copying process.
4. Using Hive, create a schema and tables based on each file's predetermined schema saved in Step 2, linking each table to a specific parquet file folder.
5. Assign Presto to use the Hive meta store and get a Presto connection string.
6. Using the Presto connection string, other analysts within the team can connect and query the data to generate insights. Moreover, the analysts can also set up a keep alive connection in a SQL data tool like Workbench, DBeaver or DataGrip whenever there's a need to test efficiency of their queries.

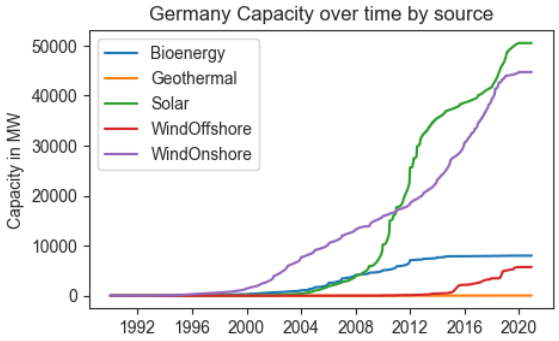
Insights & Analytics

We started by looking at the time series comparison of how the capacity uptake of the different energy sources looks like, as well as the uptake of different countries over time.

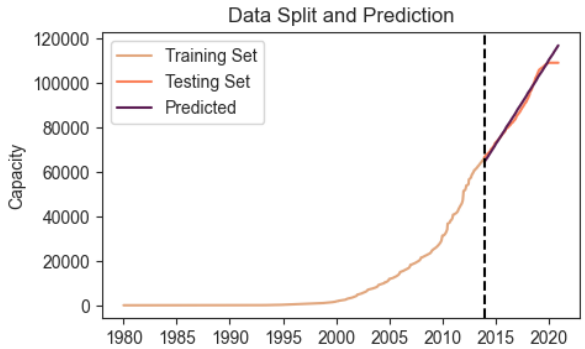


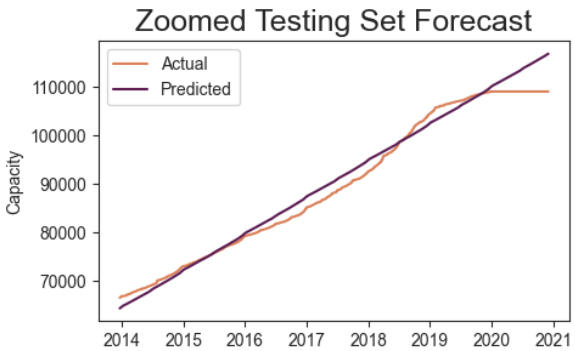


From the above we can see that Wind onshore has been the most uptaken energy source over the years followed by Solar. We also see that Hydro and Marine sources have stagnant growth over the years. From the countries, Germany has the highest update, followed by Great Britain which comprises statistics from different countries put together. We decided to focus on Germany given that it has the highest total uptake of renewable energy.

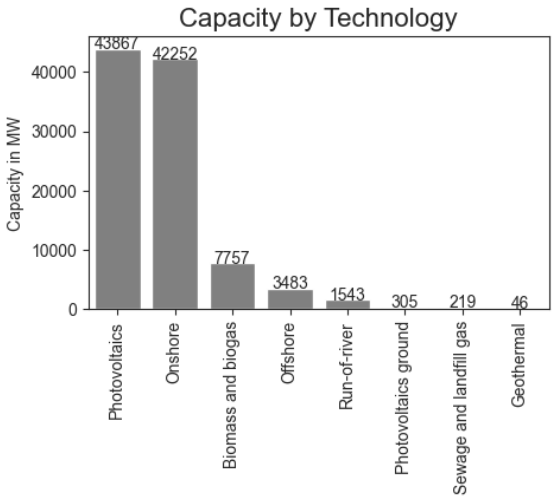


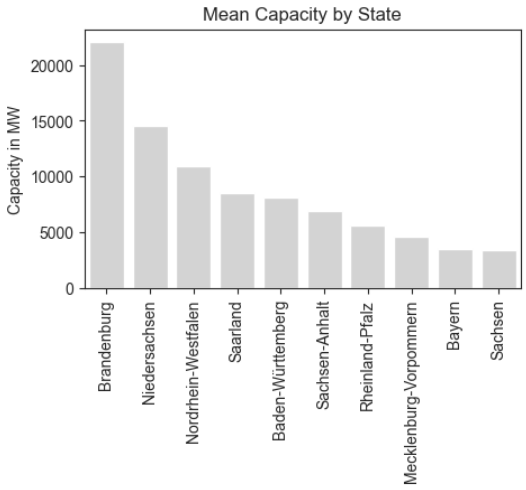
From the chart we see that in Germany, Solar is the biggest renewable energy source, followed by Wind Onshore. We can refer to these as the best two energy sources to invest in. We can use data before 2014 to come up with a growth prediction model for renewable energy consumption in Germany to predict the growth post 2014.

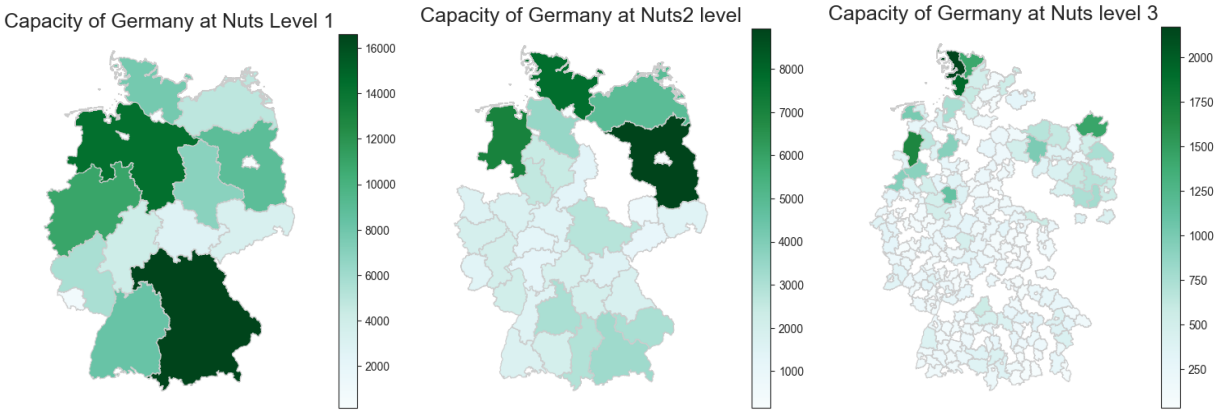




In the generation of energy, there are a couple of technologies used i.e Photovoltaics, Wind Onshore and Offshore, Biomass/Biogas, Run-of-river, Sewage and Geothermal, we can look at the total capacity of each. We can then look at capacity consumed by state, we will focus the view on the top 10 states.







From the above map charts we see that Bayern and Niedersachsen are the states (Nuts level 1) with the highest capacities. However, the two states mentioned prior have many cities and at city level (Nuts level 2), Brandenburg (Brandenburg state), and Schleswig-Holstein cities (Schleswig-Holstein state) are the top cities. At Municipality level (Nuts level 3), Nordfriesland and Dithmarschen both from Schleswig-Holstein state are the highest renewable energy consuming municipalities. The municipalities are in the North where Germany has access to the sea meaning there is massive Wind Onshore harvested using Onshore technologies, this technology ranks second meaning it should be properly established in the region. This is the ideal technology to invest in in this region.

The above approach properly breaks down the datasets used with focus on establishing a market and investment benefit, it looks at capacities at different levels including states, cities, municipalities as well as the different energy sources and the time series of the capacity uptake.

The project use cases foster investment by providing data-driven insights that reduce uncertainties, optimize decision-making, and align investments with market trends and conditions. Additionally, the project contributes to revenue generation by providing guidance in entering markets and tailoring their offerings to meet specific regional demands. The project pinpoints the regions suitable for enlargement/entering. The model used for prediction gives off a mean absolute percentage error (MAPE) of 2% which gives a good picture regarding the accuracy of the prediction.

References

[1]. Natarajan, Rahul. “AWS EMR in FS: Presto vs Hive vs Spark SQL.” Medium, 9 Mar. 2021, rahulna.medium.com/aws-emr-in-fs-presto-vs-hive-vs-spark-sql-644800c010ad.

[2]. “Presto vs Hive | Learn the Key Differences and Comparisons.” EDUCBA, 13 Oct. 2021, www.educba.com/presto-vs-hive/.